ПОСТРОЕНИЕ ЗАПРОСОВ К МАШИНЕ ПОИСКА INTERNET С ПОМОЩЬЮ ТЕЗАУРУСА

П.И.Браславский

Уральский государственный технический университет

620078, Екатеринбург, ул. Мира 34-51

pb@dpt.ustu.ru

ВВЕДЕНИЕ

По мере развития сети Internet обостряется парадокс: вероятность присутствия необходимой информации в глобальном информационном пространстве растет, а вероятность ее нахождения – уменьшается. Это происходит потому, что наполнение Сети громадно по объему, очень разнородно, быстро обновляется, плохо поддается структуризации и управлению. В этой ситуации особую актуальность приобретают исследования, направленные на повышение эффективности поиска информации в Internet.

На сегодняшний день наиболее популярное средство поиска информации в Сети – машины поиска (МП) по ключевым словам. Формулировка информационной потребности на языке запросов – наиболее сложный и трудно формализуемый этап поиска. В отличие от традиционной библиотеки, где можно обратиться за помощью к библиографу, при обращении к МП Internet пользователь оказывается "один на один" с системой.

Наше предложение состоит в том, чтобы использовать тезаурус с сильно дифференцированным набором семантических отношений в качестве основы для построения *ассистента формирования запросов* к МП Internet [1].

Сегодня тезаурусы, несмотря на богатую традицию использования в информационном поиске [6], находят лишь ограниченное применение в универсальных МП Internet. Во-первых, это объясняется тем, что задача экономии памяти и вычислительных ресурсов сегодня не так актуальна, как 30 лет назад. Во-вторых, чрезвычайно трудно построить тезаурус, который соответствовал бы тематическому разнообразию информации, индексируемой универсальной МП. Кроме того, в традиционных ИПС тезаурусы использовались в основном как средство повышения *полноты* поиска (за счет объединения близких терминов связкой "ИЛИ") [6]. Сегодняшняя проблема Internet – это информационная перегрузка (*information overload*), и как следствие – низкая *точность* поиска.

В статье формулируются подходы к использованию тезаурусов при поиске информации в Internet, описывается модель и приводится пример тезауруса с сильно дифференцированными семантическими отношениями, описывается макетная реализация метода.

ПОДХОД

На наш взгляд, тезаурус может стать эффективным инструментом формирования запросов к универсальным МП Internet и существенно повысить качество информационного поиска в специализированной тематической области при выполнении следующих условий:

- тезаурус отражает терминологию достаточно узкой научной/предметной области;
- в тезаурусе используется набор сильно дифференцированных семантических отношений;
- тезаурус независим от машины поиска (рис. 1).

Задача построения универсального тезауруса является очень трудоемкой. Построение тезауруса узкой предметной или научной области осуществить намного проще. С одной стороны, задача намного меньше по объему; с другой – в данном случае мы имеем дело не с *общеупотребительной лексикой*, а с *терминами*. Свойства терминов – системность, устойчивость и регулярность связей, отсутствие экспрессии, установка на объективность описания, – делают возможным их адекватное описание с помощью тезаурусов.

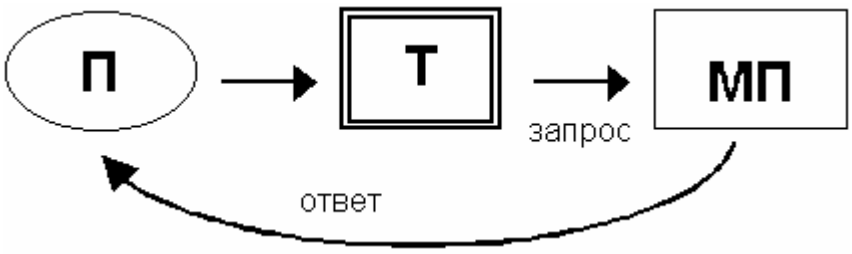

Рис. 1. Структура формирования запроса с помощью тезауруса
(П - пользователь, Т - тезаурус, МП - машина поиска)

Особенно точно описать терминологию можно при помощи тезауруса с набором *сильно дифференцированных семантических отношений* [4, 5]. Основная идея такого описания – использование не только универсальных (например, “род-вид”, “часть-целое” и т.д.), но и специфических для конкретной предметной области отношений. Таким образом, каждый *тип отношения* сам по себе несет значительную смысловую нагрузку, определяет различные аспекты семантики термина.

На основе набора семантических отношений тезауруса можно ввести понятие *стратегии поиска по тезаурусу*. Стратегия – это шаблон с указанием связки (“И”, “ИЛИ”, “НЕ”) и веса для каждого типа семантического отношения. Выбрав термин и применив к нему стратегию, мы получаем запрос, в котором выбранный (“опорный”) термин объединен со своими “соседями” в соответствии с маской-стратегией. Стратегии могут быть направлены на повышение точности или полноты поиска, выделение определенных понятийных сфер термина. Сформировав стратегию, ее можно применять последовательно к различным опорным терминам. Стратегии могут служить подсказкой начинающему пользователю, позволяют унифицировать поиск, сделать его в большей степени автоматизированным.

Средство формирования запросов на основе тезауруса позволяет тонко управлять как *полнотой*, так и *точностью* поиска.

Разнообразие, специфичность и динамика тематических интересов и информационных запросов пользователей ставит под вопрос эффективность централизованной разработки тезаурусов и расположения их на МП. Тезаурусы, отражающие терминологию различных предметных областей, могут располагаться на независимых серверах и выступать в качестве интерфейса к универсальным машинам поиска (см. рис. 1). Заметим, что Internet в такой схеме выступает не только как хранилище информации, но и как среда для коммуникации и объединения усилий разработчиков и пользователей тезаурусов.

Пользователями средства расширения запросов на основе тезауруса могли бы стать как непосредственные потребители информации (специалисты-предметники), так и те, кто занимается структурированием информационного сырья Internet: аналитические службы; специалисты, занимающиеся наполнением Internet-каталогов (*information brokers*) и т.п.

Такой инструмент может также пригодится тем, кто приступает к изучению новой области знаний, причем эффект возникает не только при изучении результатов поиска, но и при работе с самим тезаурусом – его терминами, отношениями и определениями.

МОДЕЛЬ

Важная особенность тезаурусов с набором сильно дифференцированных семантических отношений – в том, что набор отношений сам по себе несет значительную смысловую нагрузку, имеет самостоятельное значение.

Следуя [3], определим модель тезауруса как упорядоченную тройку:

$$T = \langle A, R, \Re \rangle, \qquad (1)$$

где

A – непустое множество терминов (носитель модели),

R – множество типов (символов) бинарных отношений (сигнатура модели),

$\mathfrak{R}$ – множество бинарных отношений на множестве A, причем имеется отображение множества R в множество $\mathfrak{R}$: $r \in R \Rightarrow \rho(r) \in \mathfrak{R}$ (интерпретация сигнатуры).

Из семантических ограничений следует, что все отношения $\mathfrak{R}$ нерефлексивны (термин не связан с самим собой).

Множество типов отношений должно оптимально соответствовать терминосистеме, откуда следует, что термин не может быть связан с другим более чем одним типом отношения:

$$\bigcup_{\substack{\rho_1 \neq \rho_2 \\ \{\rho_1, \rho_2\} \in \mathfrak{R}}} (\rho_1 \cap \rho_2) = \varnothing.$$

Алфавит ассистента формирования запросов содержит термины тезауруса, натуральные числа, символы связок &, |, ~, а также символы двоеточия : и скобок (,).

Слово в алфавите ассистента формирования запросов является правильным запросом, если оно удовлетворяет следующему определению:

1. Любой термин тезауруса – правильный запрос. (Если $x \in A$, то x – правильный запрос.)
2. Любой термин тезауруса с указанием веса – правильный запрос. (Если $x \in A$, $w \in \boldsymbol{N}$, то $x{:}w$ – правильный запрос.)
3. Если Q_1 и Q_2 – правильные запросы, то $(Q_1 \& Q_2)$, $(Q_1 | Q_2)$, $(Q_1 \sim Q_2)$ – правильные запросы.

Для упрощения записи можно опускать внешние скобки запроса, а также те пары скобок, без которых старшинство связок правильно определяет порядок обработки запроса.

Определим стратегию поиска по тезаурусу как упорядоченную четверку:

$$S = \langle w, R_s, \varphi_s, f_s \rangle, \qquad (2)$$

где натуральное число w – вес опорного термина;

$R_s \subseteq R$ – множество типов связей, участвующих в стратегии;

$\varphi_s : R_s \to \{\&, |, \sim\}$ – функция, ставящая в соответствие каждому типу отношения из R_s тип связки;

$f_s : R_s \to \boldsymbol{N}$ – функция, ставящая в соответствие каждому типу отношения из R_s вес.

Две стратегии S_1 и S_2 назовем совместимыми, если либо они не пересекаются по задействованным в них типам связей ($R_{S1} \cap R_{S2} = \varnothing$), либо значения функций φ_{S1}, φ_{S2} совпадают на этом пересечении ($\varphi_{S1}(x) = \varphi_{S2}(x)$, $\forall x \in R_{S1} \cap R_{S2}$).

Для двух совместимых стратегий S_1 и S_2 следующим образом можно определить объединение S:

$$S = S_1 \oplus S_2 = \langle w_1 + w_2, R_{S1} \cup R_{S2}, \varphi_s, f_s \rangle, \qquad (3)$$

где функции φ_s и f_s определяются так:

$$\varphi_s(x) = \begin{cases} \varphi_{s_1}(x), x \in R_{s_1} / R_{s_2} \\ \varphi_{s_2}(x), x \in R_{s_2} \end{cases},$$

$$f_s(x) = \begin{cases} f_{s_1}(x), x \in R_{s_1} / R_{s_2} \\ f_{s_1}(x) + f_{s_2}(x), x \in R_{s_1} \cap R_{s_2} \\ f_{s_2}(x), x \in R_{s_2} / R_{s_1} \end{cases}.$$

ПРИМЕРЫ

Пример 1. В качестве примера приведем фрагмент тезауруса, соответствующего предметной области “Диагностика и ремонт автомобилей”.

A = {автомобиль, ‘транспортное средство’, двигатель, мотор, трансмиссия, диагностика, ‘диагностический протокол’, ‘коробка передач’}

R = {род, вид, часть, целое, синоним, процесс/действие, ‘объект процесса/действия’, ‘результат процесса/действия’, ‘непосредственная связь’}

Множество отношений $\mathfrak{R}$ состоит из следующих элементов:

РОД = {(автомобиль, ‘транспортное средство’)}

ВИД = {(‘транспортное средство’, автомобиль)}

НЕПОСРЕДСТВЕННАЯ СВЯЗЬ = {(двигатель, трансмиссия), (трансмиссия, двигатель)}

ЧАСТЬ = {(автомобиль, трансмиссия), (автомобиль, двигатель), (трансмиссия, ‘коробка передач’)}

ЦЕЛОЕ = {(трансмиссия, автомобиль), (двигатель, автомобиль), (‘коробка передач’, трансмиссия)}

ПРОЦЕСС/ДЕЙСТВИЕ = {(автомобиль, диагностика), (‘диагностический протокол’, диагностика)}

РЕЗУЛЬТАТ ПРОЦЕССА/ДЕЙСТВИЯ = {(диагностика, ‘диагностический протокол’)}

ОБЪЕКТ ПРОЦЕССА/ДЕЙСТВИЯ = {(диагностика, автомобиль)}

СИНОНИМ = {(мотор, двигатель), (двигатель, мотор)}

Наглядно термины и семантические связи между ними можно представить в виде графа (рис. 2).

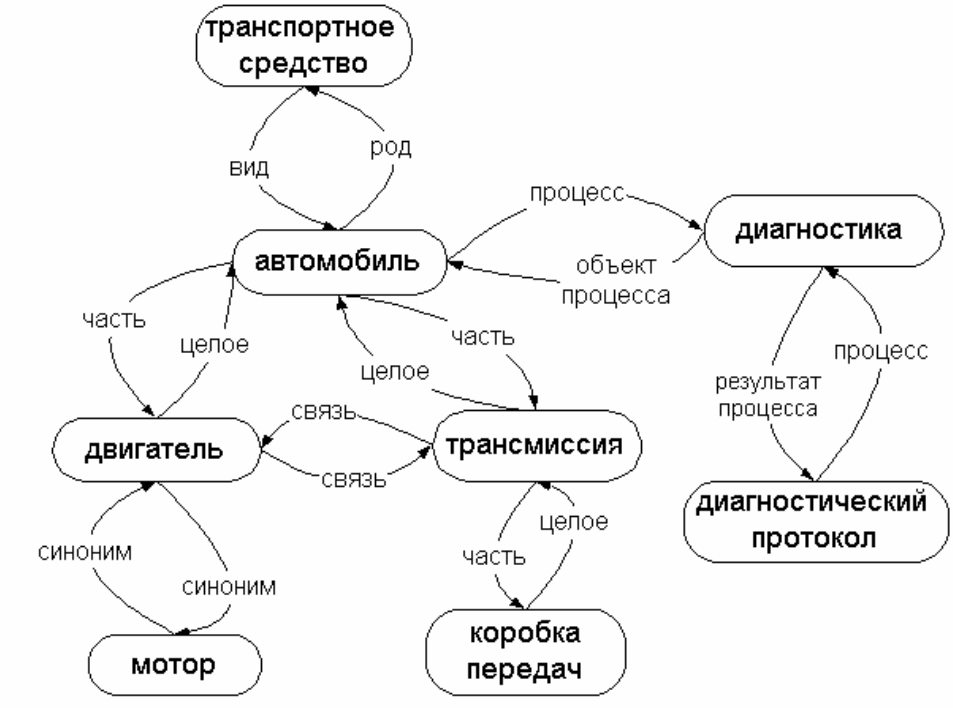

Рис. 2. Граф тезауруса из примера 1

Пример 2. Правильные запросы на основе терминов тезауруса из примера 1:

- *(автомобиль|диагностика)&трансмиссия*
- *(трансмиссия:10|‘коробка передач’:8)&(двигатель:6|мотор:4)*

Пример 3. Стратегию для тезауруса из примера 2 можно задать таблицей:

w	R_s	φ_s	f_s
10	часть	&	5
	связь	&	5
	синоним	|	8

Пример 4. Запрос, полученный в результате применения стратегии из примера 3 к опорному термину *автомобиль* тезауруса из примера 1, может выглядеть следующим образом:

автомобиль:10&двигатель:5&трансмиссия:5.

Та же стратегия, примененная к термину *двигатель*, дает следующий результат:

двигатель:10|мотор:8&трансмиссия:5.

Пример 5. Объединение стратегии из примера 3 со стратегией:

w	R_s	φ_s	f_s
2	связь	&	1
	целое	|	2

выглядит следующим образом:

w	R_s	φ_s	f_s
12	часть	&	5
	целое	|	2
	связь	&	6
	синоним	|	8

РЕАЛИЗАЦИЯ

Идеи, изложенные выше, реализованы на уровне макета в программе *ProThes Q*. Программа разработана в среде Delphi.

Интерфейс программы *ProThes Q* содержит три окна: “Термин”, “Запрос”, “Стратегия”.

Окно “Термин” предназначено для навигации по тезаурусу и выбора “опорного” термина. В окне отображаются: алфавитный список терминов; дефиниция термина (если есть) и его “соседи”.

Перемещаться по терминам тезауруса можно:

1) через поиск терминов;
2) по алфавитному списку;
3) по семантическим связям между терминами.

Окно “Стратегия” предназначено для формирования стратегий поиска. Каждому типу отношения назначается тип связки: “И”, “ИЛИ”, “НЕ” (или “пусто”, если тип отношения не участвует в стратегии). Указывается вес опорного термина и каждого типа отношения.

Окно “Запрос” (рис. 3) предназначено для работы с поисковыми запросами. Запрос формируется на основе опорного термина (выбирается в окне “Термин”) и стратегии (окно “Стратегия”). В этом окне пользователь может выбрать и скорректировать стратегию, сформировать, отредактировать и сохранить запрос. Передача запроса в программу просмотра осуществляется через буфер обмена.

Результаты обработки машиной поиска “Яндекс” запросов, сформированных при помощи программы *ProThes Q* и тезауруса по компьютерной лингвистике [2], показывают, что с помощью тезауруса можно управлять как полнотой, так и точностью поиска (таблица)[1].

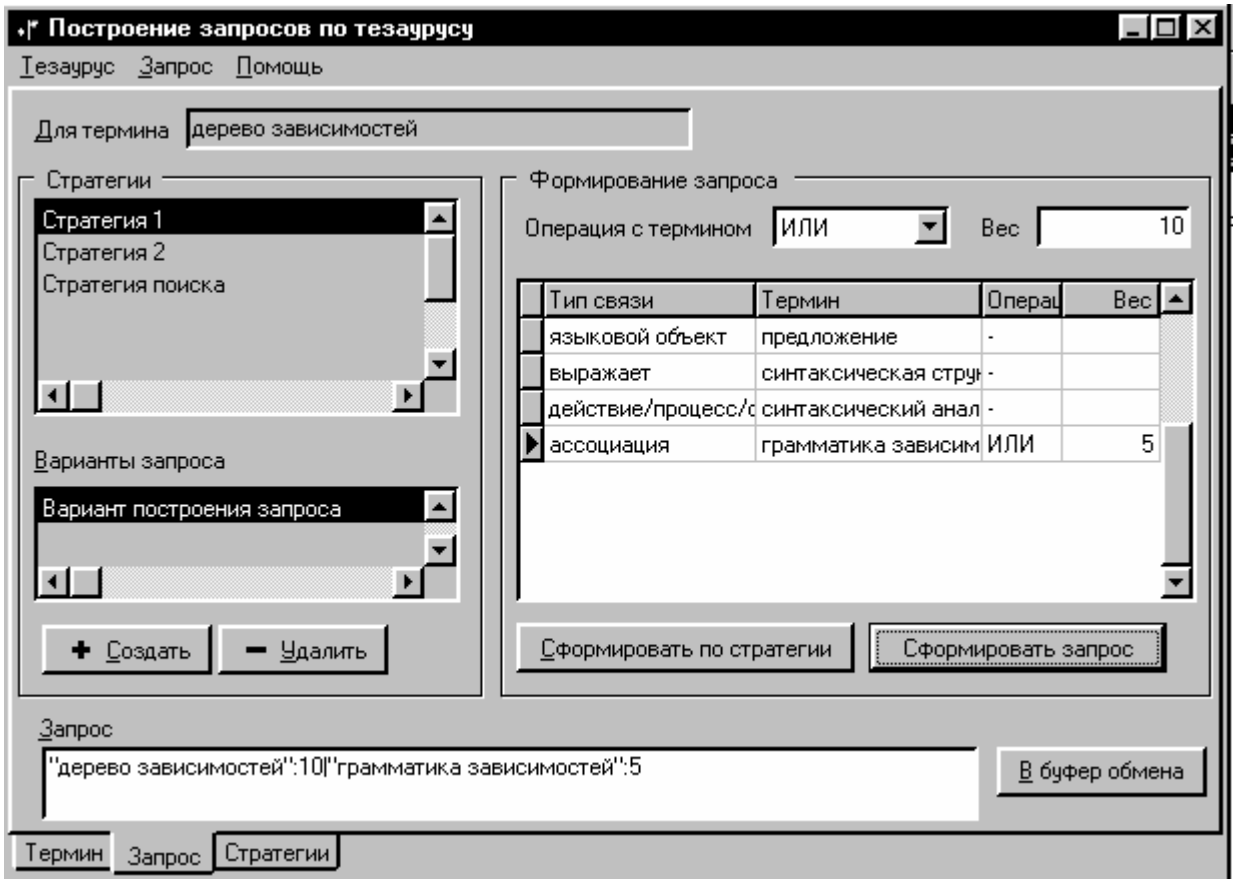

Рис. 3. Окно "Запрос" программы *ProThes Q*

Таблица

Запросы и отклики МП “Яндекс”

№	Запрос	Кол-во ссылок
1.	*“формальная грамматика”*	274
2.	*(формальная|контекстно-зависимая|контекстно-свободная| сетевая| трансформационная|зависимостей| “непосредственных составляющих”) /1 грамматика*[*)]	549
3.	*“лингвистическая трансляция”*	3
4.	*“расширенная сеть переходов”*	4
5.	*иллокуция перлокуция*	4
6.	*слово:10 предложение морфема словосочетание словоформа лексема аффикс*	68
7.	*анафора&&антецедент*	3
8.	*фрейм:10 “искусственный интеллект”*	15

[*)] Запрос был отредактирован вручную для исключения нескольких вхождений слова *грамматика*.

[1] Однако даже достаточно специфические запросы, которые формируются из терминов тезауруса, не могут гарантировать стопроцентную точность. Так, результат обработки запроса ‘*анафора && антецедент*’ содержит ссылку на документ “Цицероны филфака” – собрание смешных изречений преподавателей филологического факультета МГУ. Слова запроса встречаются в этом документе в следующих фразах: “*сермяжная теория анафоры*” и “*консеквент получили – антецедент обратно присобачили*”.

ЗАКЛЮЧЕНИЕ

Полученные результаты позволяют рассматривать тезаурусы как эффективное средство устранения дисбаланса между специфическими информационными потребностями различных групп пользователей – с одной стороны, – и универсальностью машин поиска Internet – с другой.

Важным элементом развития предложенного подхода является стандартизация формата представления тезаурусов. Формат должен отвечать требованиям открытости, переносимости и расширяемости. Выше мы уже указывали на то, что Internet должен быть не только пространством поиска информации, но и средой функционирования, разработки и поддержки тезаурусов.

Направлениями дальнейшего развития формализма тезаурусов применительно к информационному поиску могли бы стать:

- формализация процедуры формирования запроса по стратегии глубиной больше единицы (в запрос включаются не только ближайшие соседи опорного термина);
- формализация процедуры интеграция различных тезаурусов.

В заключение мы хотим поблагодарить Илью Бирюкова за помощь в создании программы *ProThes Q*.

БИБЛИОГРАФИЯ

1. Браславский П.И. Расширение поискового запроса с помощью тезауруса с сильно дифференцированными связями: Тезисы доклада рабочего совещания "Новые Интернет-технологии", Петрозаводск, КарНЦ РАН, 25-28 июня 2000 г. – http://www.krc.karelia.ru/structure/math/conf/nit/papers/paper.ru.phtml?file=braslavsky.html
2. Браславский П.И., Гольдштейн С.Л., Ткаченко Т.Я. Тезаурус как средство описания систем знаний// Научно-техническая информация. Сер.2, – 1997. – №11. – С.16-21.
3. Ершов Ю.Л., Палютин Е.А. Математическая логика: Учеб. пособие для вузов. – 2-е изд., испр. и доп. – М.: Наука, Гл. ред. физ.-мат. лит., 1987. – 336 с.
4. Никитина С.Е. Семантический анализ языка науки. (На материале лингвистики.) – М.: Наука, 1987. – 141 с.
5. Никитина С.Е. Тезаурус по теоретической и прикладной лингвистике. – М.: Наука, 1978. – 375 с.
6. Солтон Дж. Динамические библиотечно-информационные системы. – Пер. с англ. – М.: Мир, 1979. – 558 с.

Query Building to the Web Search Engines Using Thesaurus

Pavel Braslavsky
Urals State Technical University

The design of an query expansion assistant based on a thesaurus is supposed. The thesaurus reflects the terminology of an restricted area of expertise and is situated on an independent Web site. The assistant allows to build complex queries to the universal Web search engines more efficient and therefore increases the precision of the search. A model and an example of the thesaurus, a breadboard implementation of the assistant and some practical results are presented.